DOCUMENT RESUME

ED 065 547 TM 001 661

AUTHOR Planisek, S. L.; Planisek, R. J.
TITLE A Description of Fifteen Test Statistics Based Upon Optically Scanned Instructor Made Multiple-Choice Tests at Kent State University.
NOTE 11p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Data Analysis; Evaluation; *Multiple Choice Tests; *Statistical Data; *Test Construction; *Test Interpretation; Universities
IDENTIFIERS *Kent State University

ABSTRACT

Fifteen descriptive statistics were computed for 345 instructor made multiple-choice optically scanned tests for the purpose of evaluating the quality of classroom test at the university level as well as the appropriateness of the evaluating statistics. The description and analysis of the average test represented a realistic lower bound of acceptability, if the objective of testing is the assignment of grades. The results suggested specific modification of test construction practices. Insufficient range and nonsymmetric distributions suggest the need for new statistics that account for these characteristics. (Author)

A description of fifteen test statistics based upon optically scanned instructor made multiple-choice tests at Kent State University

S.L. Planisek Kent State University
R.J. Planisek Kent State University

Abstract

Fifteen descriptive statistics were computed for 345 instructor made multiple-choice optically scanned tests for the purpose of evaluating the quality of classroom tests at the university level as well as the appropriateness of the evaluating statistics. The description and analysis of the average test represented a realistic lower bound of acceptability, if the objective of testing is the assignment of grades. The results suggested specific modification of test construction practices. Insufficient range and nonsymmetric distributions suggest the need for new statistics that account for these characteristics.

A description of fifteen test statistics based upon optically scanned instructor made multiple-choice tests at Kent State University

S.L. Planisek Kent State University
R.J. Planisek Kent State University

Objectives of the inquiry:

The procedures for writing and evaluating classroom tests have become rather standardized during the past decade. With the increased availability of university computer facilities and more recently optical scanners, the application of computer methods to classroom test analysis has also become routine. It is the intent of this inquiry to raise two questions regarding the analysis of classroom tests and to provide some insight to their answers. First, based upon several typically used descriptive statistics derived from actual classroom instructor made tests, what can be said about the quality of objective multiple-choice tests at the university level? Second, do the classroom tests actually exhibit mathematical characteristics assumed by many of the statistics typically computed for the evaluation of these tests, i.e., is there justification for using the statistics typically calculated for the purpose of evaluating classroom tests?

Method and/or techniques:

All multiple choice instructor made classroom tests brought to the Examination Aids Center of Kent State University for scoring and analysis on optically scanable answer forms during a two week period selected from the middle weeks of the winter 70, spring 70, winter 71 and spring 71 academic quarters, were included in this study. The answer forms were scanned with an OpScan 100 Optical Reader and the statistics computed on Burroughs 5500 computer.

Fifteen statistics were chosen for consideration based upon their popularity among test evaluators or for their potential utility regarding answers to the two questions stated above. The typical descriptive statistics considered for each test were class size, number of items, mean, median, item easiness (or difficulty), range, standard deviation, standard error of measurement, reliability (KR-20), skewness (based on second and third moments), symmetry (mean-median), kurtosis (based on second and fourth moments), item discrimination index, item point-biserial correlation, and efficiency (range/#test items). These statistics are defined in Table 1.

A frequency distribution, mean, and standard deviation for each of these descriptive statistics were then calculated. Product-moment correlations were then computed between each pair of the fifteen variables. In addition t-tests were made among the four subsamples as a test of homogeneity.

Data sources:

The subsamples obtained in the four two week periods consisted of 78, 78, 91, and 98 tests each for a total sample of 345 classroom tests which contained approximately 21,000 items. These tests were constructed by 195 different instructors representing 28 academic departments. Although most tests were closed-book, in class, 5 response multiple-choice type tests, the sample also contained a small number of tests which were open-book, out-of-class, and/or tests which contained from 2 up to 9 response choices. Hence, the sample of tests represents a wide range of test making ability and test taking ability over a number of different content areas (refer to Table 2 and Table 4).

Results and/or conclusions:

Class size averaged 79 students per test, S.D.=71, which was significantly greater than the university average at the $p \leq .001$ level. Moreover, this large figure actually underestimated the positively skewed class size because some large classes had submitted their alternate forms as two separate tests. The symmetrically distributed test length or number of items averaged 60 questions per test, S.D.=24, with a maximum of 120 items. In addition, test length was a significant correlate with several of the variables at the $p \leq .001$ level. The average test mean was 42.6, S.D.=22; while the average test median was 42.3, S.D.=19. Knowledge of these measures of central tendency along with the average item easiness of .70, S.D.=10, an average skewness of -.34, S.D.=.60, an average symmetry (mean-median) of -.36, S.D.=.87, and an average kurtosis of .21, S.D.=3.7, raised several more questions regarding the test construction practices of college instructors (refer to Tables 3 and 5).

Furthermore, these questions were again raised by the measures of dispersion. The spread of scores or range averaged 30 score points, S.D.=12. The average test standard deviation was 6.3. Thus, the average test tended to use only half of its potential range. In addition, the average standard error of measurement was 3.0. Given the additional information that the average reliability (KR-20) was a moderate .71, S.D.=.15, the question of test validity was obvious, particularly with regard to grading practices.

An analysis of the item statistics showed the average discrimination index (based on upper and lower 27%) was .27, S.D.=.21, which indicates that at least half the items on the average test would not significantly discriminate at the .05 level. Similarly,

the point-biserial correlation coefficient for item pass/fail with total test score averaged .25, S.D.=.18, and was correlated, r=.91, with the discrimination index. Further, both of these measures of item validity were significantly correlated with the reliability at the $p \leq .001$ level. Item easiness was curvilinearly related to these statistics and produced a restricted range for several of the variables; thus, many of the poor item validities may be attributed to the large number of easy items, $\bar{E}=.70$.

The statistic which may best state the conclusion is the test efficiency, range/# of items, which averaged .50 with an S.D.=.15. The efficiency significantly correlated .79 with the discrimination index, .83 with the point-biserial correlation coefficient, and .57 with reliability at the $p \leq .001$ level. These correlations stem from the underlying assumption that additional items contribute to the test validity; however, if the additional item is too easy, $E>.73$, or too hard, $E<.27$, the contribution to validity is minimized. Thus, each additional item should increase the range a similar amount, but the correlation between the range and the number of items was .67 which again reflects the large number of easy items.

In summary, based upon the descriptive statistics derived from these instructor made tests it is the judgment of these writers that every instructor should strive to excell the average test described in this investigation. This judgement is based on the assumption that these tests were designed specifically for the purpose of assigning at least five categories of grades. About half of the tests sampled exhibited characteristics which permit valid decision making to this degree. Further, these results indicate that many

of these tests could have been reduced in length without drastically effecting their discriminatory powers simply by having eliminated those items which were too easy or too hard.

The problem of assessing the degree to which the underlying mathematical assumptions have been satisfied is not as easily answered as the first. Nonetheless, the results showed that both the average test mean and median were approximately one standard deviation from the average test length. This observation was supplemented by the measures of skewness, symmetry, and kurtosis all of which anticipated the average test easiness. The measures of dispersion and the reliability indicate a rather restricted range within which decision making is difficult. The item statistics indicate the importance of internal validity. The failure to add items which discriminate merely distorts the value of many statistics including overall test efficiency. If the objective is to validly discriminate, then perhaps the largest failure to meet underlying assumptions was found in the failure to obtain symmetric distributions.

Scientific or educational importance:

The statistical description of the average test represented a realistic lower bound of acceptability, if the objective of testing is the valid assignment of grades. Therefore, based on these results, specific modification of test construction practices are suggested. The limited range or dispersion of scores and the skewed distributions suggest the need for new statistics which are more appropriate to the nonsymmetric properties of instructor made tests.

Table 1

The Fifteen Test Statistics

Variable	Abr	Description
Class Size	N	Number of examinees for a given test
Number of Items	Q	Number of items for a given test
Mean Score	$\bar{X}$	X = raw score for student $\bar{X} = \frac{\Sigma X}{N}$
Median	Md	Middle most score
Item Easiness*	E	E = proportion of people passing the item
Range	R	R = Highest score - lowest score + 1
Standard Deviation	SD	$SD = \sqrt{\frac{\Sigma X^2}{N} - (\bar{X})^2}$
Kuder-Richardson 20	rxx	p = proportion of individuals passing item q = 1-p $rxx = \frac{Q}{Q-1}\left(SD^2 - \frac{pq}{SD^2}\right)$
Standard Error	SE	$SE = SD\sqrt{1-rxx}$
Skewness	g_1	$m_r = \frac{\Sigma(X - \bar{X})^r}{N}$ $g_1 = \frac{m_3}{m_2\sqrt{m_2}}$
Symmetry	S_y	$S_y = \bar{X} - Md$
Kurtosis	g_2	$g_2 = \frac{m_4}{m_2^2} - 3$
Item Discrimination Index*	DI	num-up = number of students in upper 27% of scores getting this item right num-low = number of students in lower 27% of scores getting this item right $DI = \frac{(num\text{-}up) - (num\text{-}low)}{.27N}$

Table 1

Variable	Abr	Description
Item Correlation Coefficient*	r	$r = \frac{(N\Sigma XY - \Sigma X\Sigma Y)}{(N\Sigma X^2 - (\Sigma X)^2)(N\Sigma Y^2 - (\Sigma Y^2))}$
Efficiency	Ef	= R/Q

*Item statistics were averaged for all items on the test and this average was used for the test statistic.

Table 2

Sample Size

Term	Number of Tests	Number of Items	Number of Students
Winter, 1970	78	4,860	5,592
Spring, 1970	78	5,138	7,332
Winter, 1971	91	5,427	7,045
Spring, 1971	98	5,564	7,330
Total	345	20,989	27,299

Table 3

Normative Statistics for Fifteen Test Statistics

Variable	Mean / Std Dev	Samples: Wtr 70	Spr 70	Wtr 71	Spr 71	Total
Class Size	$\bar{X}$	71.7	94.0	77.4	74.8	79.0
	SD	67.3	80.7	66.8	67.8	71.1
Number of Items	$\bar{X}$	62.3	65.9	59.6	56.8	60.7
	SD	19.5	26.1	24.3	24.3	24.2
Mean Score	$\bar{X}$	44.2	44.4	43.5	39.2	42.5
	SD	16.1	20.6	28.7	18.4	21.9
Median	$\bar{X}$	44.6	44.7	41.4	39.6	42.3
	SD	16.2	20.8	19.8	18.6	18.0
Mean Item Easiness	$\bar{X}$	0.70	0.67	0.68	0.69	0.69
	SD	0.09	0.11	0.08	0.09	0.10
	SD*					0.22
Range	$\bar{X}$	28.6	32.6	28.2	29.3	29.5
	SD	11.0	13.5	10.9	14.3	12.8
Standard Deviation	$\bar{X}$	6.18	6.84	6.03	6.29	6.30
	SD	2.14	2.40	2.03	2.67	2.38
KR-20	$\bar{X}$	0.70	0.74	0.72	0.72	0.72
	SD	0.16	0.11	0.12	0.18	0.15
Standard Error	$\bar{X}$	2.99	3.16	2.98	2.87	2.98
	SD	0.49	0.64	0.60	0.68	0.64
Skewness	$\bar{X}$	-0.34	-0.27	-0.29	-0.44	-0.34
	SD	0.50	0.56	0.38	0.82	0.60
Symmetry	$\bar{X}$	-0.45	-0.23	-0.30	-0.46	-0.36
	SD	0.82	0.86	0.81	0.97	0.88
Kurtosis	$\bar{X}$	0.04	0.18	-0.14	0.71	0.22
	SD	1.38	1.71	1.16	6.48	3.67
Mean Item Disc Index	$\bar{X}$	0.26	0.27	0.26	0.28	0.27
	SD	0.08	0.06	0.06	0.09	0.08
	SD*					0.21
Mean Item Corr Coef	$\bar{X}$	0.25	0.26	0.26	0.28	0.26
	SD	0.07	0.05	0.06	0.08	0.07
	SD*					0.18

Table 3

Variable	Mean Std Dev	Samples Wtr 70	Spr 70	Wtr 71	Spr 71	Total
Efficiency	$\bar{X}$	0.47	0.51	0.50	0.53	0.50
	SD	0.14	0.13	0.14	0.16	0.15

*Standard deviation based on 21,000 item statistics as opposed to the standard deviation of 345 test mean item statistics.

Table 4

t-Differences Among the Four Samples for Fifteen Test Statistics

Variable	Wtr 70 With Spr 70	Wtr 70 With Wtr 71	Wtr 70 With Spr 71	Spr 70 With Wtr 71	Spr 70 With Spr 71	Wtr 71 With Spr 71
Class Size	-1.86	-0.55	-0.30	1.45	1.71	.27
Number of Items	-0.96	0.77	1.63	1.60	2.37*	.80
Mean Score	-0.07	0.18	1.90	0.22	1.77	1.25
Item Easiness	2.06*	1.49	1.13	-0.88	-1.14	-0.34
Median	-0.00	1.19	1.87	1.06	1.68	0.64
Range	-2.01*	0.22	-0.34	2.32*	1.56	0.56
Standard Deviation	-1.78	0.49	-0.29	2.37*	1.40	-0.77
KR-20	-1.81	-0.65	-0.45	1.52	1.25	0.10
Standard Error	-1.84	0.13	1.30	1.89	2.86**	1.16
Skewness	-0.83	-0.81	0.86	0.22	1.48	1.54
Symmetry	-1.62	-1.14	0.13	0.58	1.68	1.23
Kurtosis	-0.55	0.96	-0.89	1.46	-0.70	-1.24
Disc Index	-0.80	-0.27	-1.78	0.62	-1.18	-1.74
Corr Coef	-0.92	-0.61	-2.66**	0.36	-2.09*	-2.41*
Efficiency	-1.73	-1.20	-2.64**	0.58	-1.02	-1.61

** .01 level of significance
* .05 level of significance

Table 5

Intercorrelations Among the Fifteen Test Statistics
Based on a Sample of 345 Classroom Tests

	N	Q	$\bar{X}$	Md	E	R	SD	rxx	SE	g_1	S_y	g_2	DI	r	Ef
Class Size															
Number Items	05														
Mean	01	79**													
Median	04	95**	83**												
Mean Easiness	-01	13	33**	41**											
Range	37**	67**	47**	57**	-1										
Std Deviation	17	68**	47**	55**	-15	88**									
KR-20	18	33**	26**	28**	02	67**	77**								
Std Error	11	88**	60**	71**	-15	72**	79**	46**							
Skewness	-14	-09	-14	-21*	-30**	-27**	-08	-13	10						
Symmetry	03	-12	-17	-25*	-29**	-16	-18	-20*	01	54**					
Kurtosis	26**	18	19	24*	14	35**	10	09	03	-72**	-16				
Mean Disc Index	11	40**	-35**	-46**	-23*	16	31**	60**	-13	09	-08	-17			
Mean Corr Coef	14	38**	-29**	-37**	-02	21*	29**	61**	-23*	-23*	-23*	08	91**		
Efficiency	40**	28**	-29**	-34**	-20*	44**	33**	56**	-07	-19	-07	15	79**	83**	

*Significant at the .05 level
**Significant at the .01 level
NOTE: Decimal points eliminated in each cell